Trading bias for precision: decision theory for intervals and sets

Kenneth Rice*, Thomas Lumley, Adam Szpiro

Department of Biostatistics, University of Washington

* kenrice@u.washington.edu

August 11, 2008

Abstract

Interval- and set-valued decisions are an essential part of statistical inference. Despite this, the justification behind them is often unclear, leading in practice to a great deal of confusion about exactly what is being presented. In this paper we review and attempt to unify several competing methods of interval-construction, within a formal decision-theoretic framework. The result is a new emphasis on interval-estimation as a distinct goal, and not as an afterthought to point estimation. We also see that representing intervals as trade-offs between measures of precision and bias unifies many existing approaches – as well as suggesting interpretable criteria to calibrate this trade-off. The novel statistical arguments produced allow many extensions, and we apply these to resolve several outstanding areas of disagreement between Bayesians and frequentists.

AMS 1991 subject classifications: 62C25, 62J05

Key words: Bayesian inference, interval construction, model-robustness

1 Introduction

Interval-valued decisions are fundamental in broad areas of applied statistical inference. However, while experienced statisticians will be familiar with, say, 95% confidence statements as a default approach, teachers of introductory statistics courses will also be aware that the logic behind their use is extremely subtle (Christensen, 2005). More fundamental concerns also exist, for example Savage's criticism (Savage, 1954) of the informal (but still ubiquitous) justification that intervals 'express the uncertainty' about a central point esimate. Savage (pg 398) makes it clear that, as a logical process, this development is unsatisfactory; "taking the doctrine literally, it evidently leads to endless regression, for an estimate of the accuracy of an estimate should presumably be accompanied by an estimate of its own accuracy, and so on forever".

In this paper, we review and extend decision-theoretic approaches to interval-valued inference. Section 2 reviews the previous literature on the justification of intervals as a primary goal in inference. Section 3 presents several existing approaches, with the over-arching theme of making interval decisions by trading the competing penalties of imprecision and bias. Motivated in this way, in Section 4 we present several extensions of this work, all of which are reasonably simple, but which reveal novel ties between Bayesian approaches and some existing frequentist approaches. As a whole, the paper aims to produce a formal framework for estimation under uncertainty, based directly on interval statements – and thus, in spirit, closely follows the exhortation of Casella and Berger (1987) to 'set up the interval and be done with it!'. Section 5 concludes with a discussion of how specific intervals differ in connecting statistical methods to the science at hand. While the arguments presented draw on frequentist ideas such as hypothetical replications and asymptotic approximations, and also the Bayesian incorporation of prior knowledge, we will argue that a major attraction of decision theory, as a route to understanding inference, is that the role of these abstract and subtle concepts is minimized.

2 Review

Decision theory can be applied to a broad range of inference: for a thorough book-length review, see Robert (2001). Decision-theoretic arguments for simple point-estimation and testing feature in most introductory courses on Bayesian inference, although their use is not restricted to this paradigm (Brown, 2000). The calculus is straightforward; for parameter θ and data Y, the modeling assumptions stated by the prior and likelihood allow construction of a posterior distribution $\pi(\theta|Y)$. We aim to make a decision d, which is a function of the observed data Y. In point estimation, d is real-valued, but we can view hypothesis-testing in the same framework by only considering decisions where $d \in \{0, 1\}$. A loss function $L(\theta, d)$ specifies the penalty we pay for making decision d when the truth is θ. If d is multivariate, the same calculus applies, although we note that here the loss functions map multi-dimensional decisions to one-dimensional costs. A decision is a 'Bayes rule' if it minimizes the expectation of loss under both prior and sampling uncertainty. Construction of Bayes rules is automatic; we simply minimize (with respect to d) the expected posterior loss for each set of observations Y. Straightforward examples of this calculus can be found in many introductory texts, for example (Young and Smith, 2005). Among other results, these texts show that the posterior mean of θ is the Bayes rule under squared-error loss, and that the posterior median is the Bayes rule under absolute loss.

Bayes rules enjoy several optimality properties, principally that Bayesian inference is coherent (Robins and Wasserman, 2000) – conditional on having the same prior knowledge, same data, and same loss function, then only Bayes rule cannot be 'outplayed' as decisions. Unique Bayes rules are admissible, and Bayes rules which have constant risk under some prior are additionally minimax. Also of note is the guaranteed existence of non-randomized Bayes rules, thus avoiding the use of (rhetorically weak) randomized tests. A full, formal, description of the optimality of Bayes rules is given by DeGroot (2004). Informally, for the precise 'question' asked by loss function L and the stated modeling assumptions, one can think of the Bayes rules as providing the 'best' answer, automatically.

We turn now to decision theory specifically for interval-valued decisions. The decision-theoretic calculus applies; we make a multivariate decision d calibrated by a loss function L. However, unlike point estimation, where the dimension of d and θ are identical, for interval decisions more than one component of d informs us about each element of θ: for example, the lower and upper end of an interval for a single parameter. Compared to point estimation, this many-to-one relationship between decisions and parameters means interval-decisions must consider a wider range of loss functions. Several authors have discussed loss functions for intervals; For loss functions based on considering the end-points of an interval as over- and under-estimates of univariate parameter θ, Winkler (1972) considered the intervals which are obtained as the relevant Bayes rules. For loss functions where non-coverage of the true parameter yields a step-function penalty, Cohen and Strawdermann (1973) discuss admissibility of interval-valued decisions. In relation to their earlier work on hypothesis testing (Duncan, 1975), Dixon and Duncan (1975) pursue loss-function based intervals for multiple parameters, basing their inference on t-intervals. Dixon (1976) later extended this to three-decision intervals, where deviations of the true parameter above, below or not importantly different to a null value are all treated separately. Of the more recent developments in this line of argument, the most extensive discussion is found in the papers by Casella *et al* (1993, 1994), who investigate loss functions consisting of a penalty for volume plus a step function for non-inclusion of the true parameter value. While the Bayes rule for set estimation under this loss can be calculated, it has unfortunate, 'paradoxical' operating characteristics. As noted by Berger, this behavior may be attributed to the unbounded volume component dominating the bounded step function contribution. Casella *et al* (1993) consider 'rational' loss functions where this scaling problem has been addressed, using transformations of the set volume which map it to the (0,1) range of the step function.

Robert (1994, 2001), devotes a chapter to this topic. Primarily the author discusses non-admissibility of confidence intervals, and the Stein paradox. While eschewing confidence regions whose construction "has been done in a rather off-handed manner, with no decision-theoretic justification," connections between Highest Posterior Density (HPD) intervals and classical confidence intervals are noted. The synthesis of volume, coverage

and confidence into a single loss is also considered by Robert, who notes that "this direction has not yet been treated in the literature".

2.1 Scale estimators lack Bayesian motivation

Looking ahead to Sections 3 and 4, we will supply various decision-theoretic inferences, where the fundamental goal is provision of an interval. Importantly, we will not view these intervals as secondary, 'add-ons' to point-estimation. This stance is unusual; typical frequentist approaches build intervals around chosen point-estimates $\hat{\theta}$, informally arguing that intervals 'express the uncertainty in $\hat{\theta}$', while Bayesians similarly consider that intervals 'express the posterior uncertainty in θ'. To see why this unusual stance is taken, it is helpful to consider a preliminary lemma;

Lemma 1 *Suppose we have parameters $\theta \in \mathbb{R}^p$, and wish to make a real matrix-valued, positive-definite decision R. Without restrictions on the likelihood , it holds that for any loss function $L(R, \theta)$ under which the posterior covariance is a Bayes rule for R, reporting $R \equiv 0$ is also a Bayes rule.*

To prove the lemma, first consider two point mass posteriors, with support at A and B respectively. Both posteriors have covariance equal to zero, and so for any given decision $\Xi \neq 0$, we can write

$$\begin{aligned} L_A &= L_A(\Xi, \theta) \geq L_A(0, \theta) = l_A \\ L_B &= L_B(\Xi, \theta) \geq L_B(0, \theta) = l_B. \end{aligned}$$

Now, for the discrete mixture posterior with point mass α on A and $1 - \alpha$ on B, it follows that

$$L(\Xi, \theta) = \alpha L_A + (1 - \alpha) L_B \geq \alpha l_A + (1 - \alpha) l_B = L(0, \theta).$$

Hence, whenever the posterior covariance is a Bayes rule, the rule which reports $R \equiv 0$ is also a Bayes rule. We note that it is possible to construct losses for which the posterior covariance is a Bayes rule (one trivial example is a common loss for all decisions), but

the lemma above strongly suggests that commonly-used measures of spread can not be justified from loss functions which include decisions about only spread.

Of course, Lemma 1 does not prevent the posterior variance appearing embedded in larger decisions. Secondary inference of this sort, similar in spirit to 'expressing the uncertainty' is permitted in decision theory by use of loss estimation, where point-estimation losses are expanded in some way to add on a secondary decision; see e.g. Robert and Casella (1994) . However, this approach does not resolve the fundamental criticism of Savage; arguing for this added layer of inference leads, logically, to a 'Russian doll' of uncertainty statements which are increasingly abstract. As with specification of priors in hierarchical models, the difficulty of chooising the loss functions at each level increases with its level of abstraction.

Treating the interval as the primary goal resolves this issue; precision, as interval-width, can be reflected in our utility statement, together with measures expressing the desire to have an interval near to or containing θ. No further discussion of posterior uncertainty is required, as it is reflected in the interval-width. We see that treating interval-estimation as the primary inference therefore retains the clarity and coherence provided by a single-stage decision-theoretic argument.

3 Loss functions for intervals: precision and bias

3.1 Absolute discrepancy

In the setting of a parametric model, with real-valued parameter θ, suppose we make an interval-valued decision $[a, b]$, where $a \leq b$ without loss of generality. The 'radius' of the interval, $r = (b - a)/2$ is a natural measure of precision. One similarly-natural bias term is the distance from θ to the nearest end of the interval, which we shall denote as

$$\mathcal{D}(\theta, [a, b]) = (a - \theta)^+ + (\theta - b)^+,$$

and note that this distance is zero whenever $\theta \in [a, b]$. In this way, precision and bias are measured on the same scale, in the same units, and can therefore reasonably be added.

In terms of utility, if one unit of precision has α times more utility than one unit of bias, a natural loss function is

$$L_\alpha = \alpha r + \mathcal{D}\left(\theta, [a, b]\right).$$

An equivalent determination is that our utility is unchanged if we trade one unit of precision for α units of bias. The assumption that units of r are 'bad' imposes setting known constant $\alpha > 0$, and the common assumption that bias is more important than precision similarly imposes $\alpha < 1$. Assuming only mild regularity conditions, the Bayes rule for $[a, b]$ is to report the $\alpha/2$ and $1 - \alpha_2$ quantiles of the (marginal) posterior for θ. (See, for example, Schervish (1995), pg 328). These intervals are widely used, although their justification is usually heuristic. Very attractively, quantile-based intervals automatically respect transformations of the parameter, leaving inference unaffected.

Intervals of this form have direct connections to frequentist parametric inference. By Bernstein-Von Mises Theorem (Freedman, 1999), in parametric models there is asymptotic agreement between confidence and credible intervals, and with a large sample the two will be close. (Informally, any prior information is swamped, and the asymptotically-Normal posterior matches the asymptotic frequentist distribution of the implicit point estimator at the center of the interval.)

3.2 Squared discrepancy

In the same setting as Section 3.1, keeping radius r as the measure of precision, consider trading it against a quadratic measure of bias, the squared discrepancy from θ to the center of the decision (i.e. $d = (a + b)/2$). To keep the units of precision and bias comparable, we also normalize this quadratic measure by radius r. As in Section 3.1 this allows us to calibrate the trade-off by a known constant $\gamma > 0$, and we are motivated to use loss

$$L_\gamma = \gamma r + (\theta - d)^2 / r.$$

Assuming that the posterior mean and variance are well-defined, the Bayes rule is simple; we report intervals of the form $[a, b] = [d - s, d + s]$, where

$$(d_\gamma, s_\gamma) = (\mathbb{E}\theta|Y, \gamma^{-1/2}\text{StdDev}\theta|Y).$$

(The proof minimizes a simple quadratic and is omitted). The minimized expected loss is $2\gamma^{1/2}\text{StdDev}\theta|Y$. Unlike the quantile-based intervals of Section 3.1, moment-based intervals do not respect transformations of the θ. While dramatic differences in inference due to transformations are unusual, we note that this moment-based interval is optimal only for the chosen scale of θ.

Naturally, this moment-based interval will be familiar to Bayesians - althought the derivation, and its explicit notion of optimality are novel. As in Section 3.1 familiar frequentist connections also exist. As the likelihood is approximately Normal (asymptotically) we find that setting α such that $\Phi^{-1}(1 - \alpha/2) = \gamma^{-1/2}$, confidence intervals based on Normality co-incide asymptotically with the Bayes rules of Section 3.1. Once again, the Bayesian decision-theoretic justification for these intervals requires no asymptotics, and (as we shall exploit in Section 4.4) gives a direct measure of the utility actually realized by the experiment at hand.

3.3 Step function losses

In Sections 3.1 and 3.2, we considered distances from interval-decisions to the true value, θ, where highly discrepant intervals were penalized most heavily. In situations where the parameter of interest can be measured in 'real-world' units (such as lengths, times, and weights) then these utility statements seem appropriate and natural. By contrast, where the parameter of interest has only discrete values intervals do not seem a helpful approach. However, we are careful to distinguish this from settings with a continuous parameter where the utility is discrete, in which (as we see below) intervals can be formed.

If our concern lies solely in making the 'right' binary decision, with no regard to the available precision, then a natural utility is an indicator function that we report the

correct statement. For point-valued decisions, a natural loss function is an indicator (formally, a Kronecker-δ function) that decision d reports the true θ, exactly. Under such a loss, the Bayes rule is to report the posterior mode. Where this decision is a set, we are 'right' when the decision set and truth are concordant, e.g. reporting decision $D = \mathbb{R}^+$ is 'right' if and only if $\theta > 0$. This motivates the utility

$$L(D, \theta) = |D| + c\mathbf{1}_{\theta \notin D},$$

where $|D|$ denotes the size of D with regard to Lebesgue measure on the θ-space – i.e. not with regard to the posterior. Writing the posterior density as $f(\theta|Y)$, the Bayes rule in this case report the set of points for which $f(\theta|Y) > c^{-1}$ (Schervish 1995, pg 329). This is a Highest Posterior Density (HPD) interval, motivated as an extreme case of trading precision (size of D) for bias (getting the 'right' answer, or not). Unlike the quantile- and moment-based intervals above, HPD intervals can be disjoint, when the posterior is multimodal. However, as HPD intervals follow moment-based intervals in not respecting transformations of θ, their optimality holds only if the scale of θ is made explicit.

As a trade-off between precision and bias, the step loss function above has unattractive features. First, for continuous θ, the uniform penalty for all incorrect values of θ seems implausible in scientific inference – where knowing the actual value of θ would usually be the ultimate goal, and getting any degree closer is helpful. (This is distinct from other fields, such as drug-licensing and legal judgements, where after a terminal binary decision, the 'truth' becomes irrelevant.) A second, general, concern is the choice of c; this has units which are the reciprocal of posterior density, and is therefore not easily interpreted; for example, its meaning changes with sample size. Moreover, the frequentist performance of decisions based on losses have unfortunate, 'paradoxical' operating characteristics, as noted by Casella *et al* (1993) and Berger (1993). This may be attributed to the unbounded precision component dominating the bounded step function, penalizing for bias. Casella *et al* (1993) consider 'rational' loss functions where this scaling problem has been addressed, using transformations of the set volume which map it to the (0,1) range of the step function, but these have not been generally adopted.

The decision-theoretic approach given here is not the usual Bayesian motivation of HPD intervals, where one reports the smallest set providing a particular percentage of posterior support – almost always 95% percent, in practice. This is not a trade-off of bias versus precision; it fixes the precision and proceeds to minimize the bias. While this minimization does have a decision-theoretic justification, its imposition of say, 95% credibility is a purely statistical step, with no direct link to scientifically-measurable quantities. Of course, for unimodal posteriors, the HPD intervals will typically be close to moment- and quantile-based intervals, and, more generally, asymptotic Normality of the posterior provides the same links to frequentist behavior as seen in Sections 3.1 and 3.2.

4 Precision and bias: extensions

4.1 Reconciling Bayesian and frequentist testing, and Lindley's paradox

We now generalize the loss functions of Section 3.1 to include testing. Suppose, in addition to interval (a, b), we were additionally interested in making a binary decision h. Just as we have penalized both precision and bias above, we can penalize discordance between binary and interval decisions. Specifically, if there exists some 'null' parameter value θ_0, reporting $h = 1$ for $\theta_0 \in [a, b]$ is discordant, and similarly $h = 0$ for $\theta_0 \notin [a, b]$. The distance of θ_0 from $[a, b]$ measures this discrepancy on the same scale as the precision and bias, motivating the following loss function;

$$L_\alpha = \alpha r + \mathcal{D}\left(\theta, [a, b]\right) + h\mathcal{D}\left(\theta_0, [a, b]\right) + (1 - h)\mathcal{D}\left(\theta_0, [a, b]^C\right),$$

where the final term denotes the minimum of the distances from θ_0 to the end points of $[a, b]$. The Bayes rule is a simple extension of Section 3.1; both discordancy measures are minimized, to zero, when we report the usual quantile-based interval for $[a, b]$ and a simple indicator function that $\theta_0 \notin [a, b]$ for h. (Although not pursued here, the addition of a binary decision extends the moment-based intervals of Section 3.2 in an

exactly similar manner.)

The loss here merely formalizes the argument of Wald , whose desideratum was that tests and intervals should agree. It is therefore unsurprising that we derive a Wald-type test, but the consequences are not trivial. Appealing to Bernstein-Von Mises' large-sample agreement between confidence and credible intervals in parametric models, it follows that, to a good approximation, this Bayes rule returns $h = 1$ whenever θ_0 is excluded from default confidence intervals. Re-arranging this property, we can define

$$p^* = \min\{\alpha | \theta_0 \notin [a_\alpha, b_\alpha]\}$$

and note that we declare $h = 1$ — 'rejecting the null' — if and only if $p^* < \alpha$. Strictly speaking, the p^*-value is not a Bayes rule, but it is a sufficient summary of the posterior to allow computation of the (binary) Bayes rule. In this way it is a natural analog to the two-sided p-value in the Neyman-Pearson framework; p is not itself a decision, but is a sufficient summary of the data to permit calculating the binary, hypothesis-testing decision. Consequently, decision-theoretic Bayesians can accept inference based on classical two-sided p-values, at least as an approximate 'answer' to the 'question' posed by L_α above. Of course, this appears to be at odds with Berger and Sellke's famous result (Berger and Sellke, 1987), that two-sided p-values and Bayesian measures of evidence are "irreconcilable", but the Berger and Sellke framework restricts attention to a single binary parameter. Like the step-function losses of Section 3.3, this binary parameter permits only the trading-off of costs for Type I and Type II errors, and disregards precision, bias, and their combined discordance with the null value.

Having shown that there are at least two distinct Bayesian ways to summarize the data for testing, it seems appropriate to point out the frequentist equivalent. While the classical Wald-style frequentist approach uses p-values, more recently frequentists have used two-part models for testing. For example, Efron's 'local false discovery rate' (Efron, 2008) makes testing decisions by estimating the long-run frequency of erroneous decisions, conditional on future data looking like current observations. Bayesians will recognize this as the posterior probability of the null, one part of the data-dependent term in the Bayes Factor.

These different ways of testing are laid out here to demystify Lindley's famous paradox (Lindley, 1957), where default Bayesian and frequentist tests lead to dramatically different conclusions — a situation described recently as "most troubling for statistics and science" (Berger, 2003). Our argument suggests that, while the distinction is real, it is essentially one of interval versus binary decisions, and is not fundamentally Bayesian versus frequentist.

4.2 Bayesian Bonferroni correction

Consider the generalization of Section 3.1 to a multivariate parametric problem, where the interest lies in making m interval-valued decisions. For interval decisions $[a_1, b_1]$,$[a_2, b_2]$, ..., $[a_m, b_m]$, a naïve extension of the quantile-based intervals is obtained by using loss

$$\sum_{j=1}^{m} \alpha \frac{b_j - a_j}{2} + \sum_{j=1}^{m} \mathcal{D}(\theta_j, [a_j, b_j]). \tag{1}$$

The Bayes rule is established by minimizing each summand with respect to its constituent $[a_j, b_j]$, after taking the expectation with respect to the posterior, and therefore to report the set of $(\alpha/2, 1 - \alpha/2)$ quantiles, for each $\theta_1, \theta_2, ...\theta_m$ respectively. The derivation is simple, but is made explicit here to contradict the prevalent misconception that 'Bayesians automatically adjust for multiple comparisons' (Mueller et al., 2007). While this statement has some justification for Bayesian posterior probabilities, where the goal is model selection (Moreno and Girón, 2006), such statements are invalid for interval-based inference, and therefore invalid in general. Without any ambiguity, we see that adjustment for multiplicities is not an automatic consequence of basing inference on the posterior distribution. However, in interval-based inference, Bayesians can allow for multiplicities through their choice of utility, as we now explore.

We can consider (1) in terms of its precision-bias trade-offs; the sum of 'precisions' is traded, at rate α against the sum of 'biases'. However, in a multiple-parameter setting, a more natural trade-off may be between the sum of the bias terms, which increases if m increases, against the average precision, which should remain roughly constant. This

motivates using loss

$$L_\alpha^\diamondsuit = \frac{\alpha}{m}\sum_{j=1}^{m}\frac{b_j - a_j}{2} + \sum_{j=1}^{m}\mathcal{D}(\theta_j, [a_j, b_j]).$$

By a substitution in the result above, we see that the Bayes rule is to report the set of $(\alpha/2m, 1-\alpha/2m)$ quantile intervals for each θ_j. Appending the binary decisions as in Section 4.1, each 'corrected' p_j^*-value is m times larger than the naïve approach would give.

This result will be familiar as Bonferroni correction, to either intervals or Wald-style p^*-values. Up to the agreement of credible and confidence intervals, it is seen to be an optimal and natural Bayesian procedure. Usually, Bonferroni correction is seen as a crude way to control the (frequentist) probabilitiy of making any Type I error; the Family-Wise Error Rate (FWER). As such, it is commonly refered to as 'conservative'. In fact two distinct types of conservatism are at play; first, the FWER is a conservative basis for decision-making, worrying about making any Type I errors, regardless of m. The Bayesian analog is no less conservative, although the chosen utility at least makes it explicit how precision and bias are being weighted. Secondly, Bonferroni's control of realized FWER is 'conservative', in that the actual level is below the nominal α. For correlated test statistics, and large m, Bonferroni's frequentist conservatism can be substantial (see e.g. Bland (2000)). By contrast, for any m the Bayesian analog is exactly optimal for the loss $L_\alpha^\diamondsuit$, whatever the correlation between the different θ_j in the posterior, as only the quantiles of the marginal distribution of each θ_j are reported.

4.3 Bayesian FDR correction

In Section 4.2, we noted that a trade-off is made between average precision and total bias. A minor modification to this is to trade-off average precision and positive-average bias, that is, the average of all biases which are strictly positive. For multivariate parametric inference, we therefore consider the utility

$$L_\alpha^\triangle = \frac{\alpha}{m}\sum_{j=1}^{m}\frac{b_j - a_j}{2} + \frac{1}{N \vee 1}\sum_{j=1}^{m}\mathcal{D}(\theta_j, [a_j, b_j]),$$

where $N = \#\{\mathcal{D}(\theta_j, [a_j, b_j]) > 0\}$. As in Section 4.2, we can find Bayes rule intervals minimizing the expected loss, although they may not be available in a closed form. Nevertheless, following section 4.1, we can also append binary decisions, denoting whether the null values θ_{0j} are within these intervals. Given the similarity of loss above to section (1), it seems natural that the p_j^* generated for 'naïve' decisions will be a useful numerical first step for deciding if the appended decisions here are 0 or 1.

In Appendix 1, we show that for posteriors under which each θ_j is independent, just such a short-cut exists. Formally, if we order the p^*-values, so $p^*_{[1]} \leq p^*_{[2]} \leq ... \leq p^*_{[m]}$, then an optimal way to find 'rejected' null values, is to search for the maximum p_j^*, below Simes' line, i.e.

$$\max \{j : p^*_{[j]} \leq \alpha j/m\}.$$

This is the well-known Benjamini-Hochberg (BH) algorithm (Benjamini and Hochberg, 1995), applied to p^*-values. Analogously to the discussion of p^*-values, we make no claims that the BH algorithm is making optimal Bayesian decisions. Indeed, it is known that BH's decisions can have unfortunate operating characteristics, albeit in rather specialized circumstances. (For some simple and compelling examples, see Fearn (2007)). However, the explosion in FDR and related work since the seminal BH paper suggests, empirically, that these approximate Bayesian decisions are extremely useful in practice; investigating utilities similar to $L_\alpha^\triangle$ therefore provides substantial scope for refining and improving the existing methods.

4.4 Reconciling Bayesian and Frequentist sample size calculations

In both Sections 3.1 and 3.2, the minimized expected losses combine precision and bias in a measure of overall utility. As sample size increases, we would expect both precision and bias to improve, and so the realized utility should also improve. Choosing n large enough so that the expected loss is below a preset threshold therefore defines a sample-size calculation. Moreover, this sample-size calculation is naturally concordant with our

inference, based on exactly the same loss function.

As noted in Section 3.2, moment-based intervals are derived from a loss with an optimized expected value directly proportional to the posterior standard deviation. For very general models, this quantity is proportional to $\sqrt{n}$, and so linear changes to the required utility threshold will equate to quadratic changes in required sample size. As a specific illustration, consider use of L_γ from Section 3.2 with independent identically distributed $Y_i \sim N(\theta, \sigma^2)$, for $1 \leq i \leq n$ and where σ^2 is known. We assume the standard conjugate prior, with $\theta \sim N(m, v)$, parameterized with $v = \sigma^2/k$, so k can be informally interpreted as 'prior sample size'. The posterior for θ is thus $N(\bar{Y}\frac{n}{k+n} + m\frac{k}{n+k}, \sigma^2\frac{1}{n+k})$. Scaling L_γ by a constant, we can express the loss as

$$Z_{1-\alpha/2}\frac{Z_{1-\alpha/2}+Z_{1-\beta}}{2}\left(\frac{a-\theta}{2\sigma}+\frac{b-\theta}{2\sigma}\right)^2\frac{2\sigma}{b-a}+Z_{1-\alpha/2}^{-1}\frac{Z_{1-\alpha/2}+Z_{1-\beta}}{2}\frac{b-a}{2\sigma}.$$

The 'basic' cost-per unit of loss is the average of fixed constants $Z_{1-\alpha/2}$ and $Z_{1-\beta}$, but to reflect conservatism about bias, the cost of precision down-weighted by a factor of $Z_{1-\alpha/2}^{-1}$, while the bias term's cost is upweighted by the same factor. Using the Bayes rule decision, the expected posterior loss is

$$\mathbb{E}L^\dagger(d_\gamma, \theta)|Y = 2\sqrt{\gamma}\sigma/\sqrt{n+k}.$$

If our threshold for this quantity is Δ, we should therefore insist on a sample size of

$$n + k \geq \frac{\sigma^2}{\Delta^2}(Z_{1-\alpha/2} + Z_{1-\beta})^2.$$

Taking the 'non-informative' limit $k \to 0$ then we recover the familiar frequentist sample size calculation for the two sample problem (see e.g. van Belle et al (2004) pg 136);

$$n \geq \frac{\sigma^2}{\Delta^2}(Z_{1-\alpha/2} + Z_{1-\beta})^2,$$

for standard Normal quantiles $Z_{1-\alpha/2}$ and $Z_{1-\beta}$, where α and β determine the Type I and II error rates.

As in Section 4.1, our interval-based decision-theoretic derivation goes against a substantial Bayesian literature criticizing the default frequentist calculation. (A review, and partial reconcilliation is provided by Inoue et al (2005), but this requires weighting

of binary, hypothesis-testing decisions.) Typical Bayesian approaches are instead based on the maximized expected utility (MEU) argument of Lindley (1997). Here, one appends a cost term (usually linear in n) to a loss function concerned with point-estimation of θ, such as squared-error loss. Treating n as a decision, the Bayes rule then provides the unique optimal trade-off between cost and a measure of bias. This trade-off between bias and cost is awkward in two simple ways: First, one is forced to trade-off units on totally different scales. Second, MEU forces the acceptence of a uniquely 'best' n, when in practice if there was more funding available, to obtain better statistical accuracy, everyone would prefer bigger samples.

4.5 Bayesian sandwich estimates

We now generalize the approach of Section 3.2 to the case of multivariate $p \times 1$ vector θ, in a non-parametric setting. Our goal is a decision ellipsoid, characterized by $p \times 1$ vector d (center) and positive definite $p \times p$ matrix R^2 (radii and eccentricity).

As a first step, we consider a parametric setting. Generalizing the loss of Section 3.2, we can characterize 'precision' by the sum of the longest orthogonal axes of the ellipsoid, and 'bias' by sum of squared distances from θ to the center of the ellipsoid, normalized to maintain dimensional comparability. Trading off as in Section 3.2, our loss function is

$$L_\gamma = \gamma tr(R) + (\theta - d)^T R^{-1} (\theta - d).$$

As shown in Appendix 2, the Bayes rule is to report

$$d = \mathbb{E}\theta|Y, \;\; R^2 = \gamma^{-1}\mathrm{Cov}\theta|Y,$$

where we assume these first two moments are well-defined.

Before extending this result in a model-robust manner, we note that non-parametric measures of precision and bias requires some care. The traditional Bayesian paradigm has *prima facie* difficulties with non-parametric inference, because parametric likelihoods are the link between prior and posterior. Typically, inference about parameters made under specific model assumptions will not have an automatic interpretation under other

model assumptions. However, if inference is instead made on quantities which can be defined equivalently under broad classes of models, this difficulty does not arise.

Suppose we are in the setting of multivariate linear regression, with a fixed $n \times p$ design matrix X, and a $n \times 1$ vector of independent outcomes Y. Our goal remains a multivariate ellipsoid decision, centered at d and with radii/eccentricity matrix R. As in the parametric framework, $tr(R)$ is a natural measure of ellipsoid precision. A measure of 'bias' that is comparable across models is the expected squared error loss around Xd, where the expectation is taken with respect to replications from the (as yet unspecified) data-generating model. Now, this 'bias' around replicated data has n dimensions, not p; by dimensional considerations, its trade-off against precision requires a projection of the replicated data to $\mathbb{R}^p$. A natural projection is a generalized inverse of X, such as $\Pi(X) = (X^T X)^{-1} X^T$.

Motivated in this way, consider the loss function

$$\gamma tr(R) + \mathbb{E}_{Y^*|\theta}(Y^* - Xd)^T \Pi(X)^T R^{-1} \Pi(X)(Y^* - Xd), \tag{2}$$

where $\mathbb{E}_{Y^*|\theta}$ denotes expectation with respect to the (as yet unspecified) data-generating mechanism, using its parameters θ. The loss is a function of the decision and θ alone. As shown in Appendix 3, the Bayes rule is to report

$$\begin{aligned} d &= \Pi(X)(E)Y^* \\ R^2 &= \gamma^{-1} \Pi(X) \mathcal{E}(Y^* - \mathcal{E}Y)(Y^* - \mathcal{E}Y)^T \Pi(X)^T, \end{aligned}$$

where $\mathcal{E}$ denotes expectation with respect to the posterior predictive distribution, i.e. generating new datasets under the updated (but still unspecified) data-generating model. When this model assumes that the elements of Y^* are independent, the Bayes rule for R^2 simplifies to

$$R^2 = \gamma^{-1} \Pi(X) \text{diag}\{\text{Var} Y_i^*\} \Pi(X)^T.$$

These formulations parallel the popular 'sandwich' estimates of covariance, due to Huber, Eicker, and White (Huber, 1967). The Bayes rules depends only on the first and second moments of $\Pi(X)Y^*$. In particular, we obtain the same inference for any and all models

which are sufficiently flexible to reproduce first and second moments of $\Pi(X)Y$, so gaining robustness to choice among this class of models. (In fact, the sandwich approach may be viewed as an attempt to restrict frequentist inference to properties of only first and second moments of $\Pi(X)Y$ (White, 1980); in this light the Bayesian connection is natural.)

Another derivation of sandwich intervals is via construction of θ as a straight line fit to the posterior from a highly-flexible model, followed by use of L_γ above. The equivalence of θ across broad classes of models is less explictit than that defined in (2), although this approach is closer to existing 'Non-parametric Bayes' methods. Through specific choices of θ, this construction can be used for fixed or random X (Szpiro et al., 2007).

Sandwich-based intervals are the default approach in modern frequentist analysis (Liang and Zeger, 1986; Zeger and Liang, 1986; Zeger et al., 1985), and enjoy the considerable benefit of asymptotically-correct coverage under only mild regularity conditions. A full discussion of the consequent freedom from model-validity that this provides is given in Section 5.2, but we note that establishing a Bayesian interpretation of this approach has been an open problem for some time.

5 Discussion

5.1 Comparisons with existing frequentist and Bayesian methods

Our developments of intervals has primarily been within the Bayesian paradigm, although strong connections have been made to frequentist ideas throughout Section 4. While these results may bring Bayesian and frequentist agreement, we feel the in-built optimality of our approach is more readily-accessible than current Bayesian and frequentist practice.

The default frequentist method is to find intervals which, under hypothetical replications of the experiments, and appealing to asymptotic approximations, have acceptably

accurate coverage of the true value. The 'not too wide' criterion is met, informally, by using asymptotically efficient estimates, where these are known. If the calculated standard errors are good approximations to the true standard error of the efficient estimator, then we obtain intervals with pre-specified coverage which are of minimal average width, averaging over replications of the experiment. However, these intervals' frequentist 'optimality' is somewhat removed from the actual data at hand (Christensen, 2005). In many fields, this disconnect encourages naïve users to examine several different confidence intervals, all of which may be valid in the frequentist sense. In an essentially *ad hoc* step, the practitioner then reports whichever is deemed to 'work best'. By comparison, the decision-theoretic approach encourages the user to consider, *a priori*, both what is known and what we want to know. Optimal inference then follows directly.

Optimality is also remote from default Bayesian derivations of intervals. While HPD intervals are derived from a length-optimizing argument, they are employed rarely, typically for multi-modal posteriors. In practice 95% quantile-based intervals are preferred. These are commonly justified only as a 'reasonable summary' of the posterior, and the rationale for preferring them to HPD intervals is not stated.

Where data-exploration is the goal, any frequentist or Bayesian method of interval-construction may be appropriate. However, for formal inference on a pre-specified quantity of interest, more rigor is demanded.

5.2 Model-checking versus utility-specification

The logic of the decision-theoretic approach is quite straightforward; we specify a question of scientific interest, and construct an optimal answer based on the data. While this approach may be familiar in point-estimation problems, in this paper we have shown that the same approach generalizes to decisions which are intervals, or sets. We emphasize that the approach is confirmatory, not exploratory; although the inference relies on the data, no *post hoc* changes are made to the question being addressed.

Importantly, model-checking need play no part in this inferential process. In the Sec-

tions on parametric work, we condition on belief in the postulated model – a typical assumption in parametric approaches. However in Section 4.5, while this conditioning occurs it is much less important, and we gain model-robustness in the same spirit as Huber, Eicker and White. Having shown that identical inference is obtained over a broad class of assumptions, if we accept the validity of any one of these assumptions, logically we can accept the consequent inference.

This approach diverges substantially from practices familiar from both Bayesian and frequentist parametric inference. Here, to allay concerns over the strength of the assumption that the parametric model is correct, informal algorithms exist to iteratively check and enhance the 'assumed' model; see, for example, McCullagh and Nelder(McCullagh and Nelder, 1999), pg 392. While this framework seems uncontroversial for *ad hoc* data exploration, it is known to be unsatisfactory for formal, confirmatory analysis, and open to substantial misuse. At a simple level, well-intentioned but naïve use of p-values uncorrected for model-checking can result in aberrant Type I error rates (Caudill, 1988; Forsberg and Kristiansson, 1999). More seriously, in clinical trials of new drugs, this informality of this approach allows ample opportunity for deliberate obfuscation. Consequently, FDA guidelines effectively rule out model-checking in such work (Senn, 2000). Attempts have been made to formalize the cycle of exploration/confirmation, but these are not yet generally accepted.

In contrast with common presentations of model-robustness as a 'clever trick', we feel that its justification from a decision-theoretic point of view follows as a logical consquence of simply asking a model-robust question. In particular, this development is distinct from the justification that we are 'robustifying the MLE,' an argument comprehensively denounced by Freedman (Freedman, 2006). If we ask a model-robust question, there may be no need for model-checking, logically or practically. With a small sample, a frequentist might be concerned about the accuracy of asymptotic approximations, while a Bayesian might be concerned that their prior had not been swamped by information from the data. However, in large samples, using well-understood models, these concerns can quite reasonably be ignored.

This divergence from received statistical wisdom is not intended to be taken lightly, although we do note that 'sandwich' approaches, without model-checking, are already the default in some applied fields, often via the Stata software. However, we are enthusiastic that its decision-theoretic justification of model-robust approaches will allow statisticians more generally to be comfortable using method which allows them to concentrate directly on the science at hand. In the model-robust framework, the statistician's challenge, instead of formally modelling the 'truth', becomes simply to formalize the scientific question of interest; in this way, statistical inference can follow the science, and not *vice versa*, as can happen with other methods. Particularly in situations where no parametric model offers a plausible 'truth' for the data, we feel the framework is not only epistemologically appealing, but also offers a better practical use of statisticians' talents than a quixotic search for 'better model fit'.

6 Appendices

6.1 Appendix 1

The expected loss is

$$\frac{\alpha}{m}\sum_{j=1}^{m}\frac{b_j-a_j}{2}+\mathbb{E}\frac{1}{N\vee 1}\sum_{j=1}^{m}\mathcal{D}(\theta_j,[a_j,b_j]),$$

where the expectation is with respect to the posterior. Now, writing $\tilde{N}_j$ as the number of positive bias terms except the jth, i.e. $\#\{\mathcal{D}(\theta_k,[a_k,b_k])>0,\, k\neq j\}$, this is

$$\frac{\alpha}{m}\sum_{j=1}^{m}\frac{b_j-a_j}{2}+\sum_{j=1}^{m}\mathbb{E}\frac{\mathcal{D}(\theta_j,[a_j,b_j])}{1+\tilde{N}_j}.$$

In the jth summand only θ_j appears only in the numerator; not θ terms appear in the denominator. Hence for posteriors in which all parameters are independent, we can write

$$\frac{\alpha}{m}\sum_{j=1}^{m}\frac{b_j-a_j}{2}+\sum_{j=1}^{m}\mathbb{E}\left(\frac{1}{1+\tilde{N}_j}\right)\mathbb{E}\mathcal{D}(\theta_j,[a_j,b_j])$$

Interim to finding the Bayes rule, we take derivatives with respect to a_j and b_j. Writing F_j as the marginal posterior distribution of θ_j, and f_j as its density function, setting

these derivatives to zero leads to

$$\begin{aligned} -\alpha/2m + \mathbb{E}\left(\frac{1}{1+\tilde{N}_j}\right) F_j(a_j) + \mathbb{E}\mathcal{D}(\theta_j,[a_j,b_j]) \sum_{k\neq j} f_j(a_j)\mathbb{E}\left(\frac{1}{2+\tilde{N}_{jk}} - \frac{1}{1+\tilde{N}_{jk}}\right) &= 0 \\ \alpha/2m - \mathbb{E}\left(\frac{1}{1+\tilde{N}_j}\right)(1-F_j(b_j)) - \mathbb{E}\mathcal{D}(\theta_j,[a_j,b_j]) \sum_{k\neq j} f_j(b_j)\mathbb{E}\left(\frac{1}{2+\tilde{N}_{jk}} - \frac{1}{1+\tilde{N}_{jk}}\right) &= 0, \end{aligned}$$

where $\tilde{N}_{jk}$ extends the notation above, representing the number of positive bias terms excluding the jth and kth. In both expressions the summand with $j \neq k$ behaves as $\tilde{N}_{jk}^{-2}$, and so for large numbers of positive bias terms, an approximation to the Bayes rule can be found be setting this term to zero. (This simple approximation is utilized below) Having made this approximation, trivial re-arrangement shows leads to setting $F_j(a_j) = 1-F_j(b_j)$, i.e. reporting quantile-based intervals, symmetric about the posterior median for θ_j, exactly as for $L_\alpha^\diamondsuit$. However, even with the approximation step, the required quantiles $F_j(a_j)$ are not completely trivial to calculate, being defined as the solution of the system of m related equations;

$$F_j(a_j) = \frac{\alpha}{2m\mathbb{E}\frac{1}{1+\tilde{N}_j}}, 1 \leq j \leq m.$$

Minimizing the exact posterior loss, in order to compute the exact Bayes rule $d_\alpha^\triangle$ appears still more challenging.

This numerically challenging work may be excessive, however, if we follow the line of argument in Section 4.1, and only require a summary of the $\{[a_\alpha^\triangle, b_\alpha^\triangle]\}$, i.e. the set of exact Bayes rules indexed by α. Given a vector of null values $\theta_{0,j}$, a natural extension of the p^*-value is for a specified α, to ask whether $\theta_{0,j} \in [a_{\alpha,j}^\triangle, b_{\alpha,j}^\triangle]$, i.e. whether the null value lies within the exact Bayes rule interval. As we see below, this property may be determined without calculating the exact Bayes rule.

Following (6.1) we can define the p^*-value for each parameter as

$$p_j^* = F_j(\theta_{0j})$$

Without loss of generality, we assume that all null values lie below their respective posterior medians, so that $F_j(\theta_{0,j}) \leq 1/2$. Now, let us assume for some j that $\theta_{0,j+1} \in$

$[a^{\triangle}_{\alpha,j+1}, b^{\triangle}_{\alpha,j+1}]$, and that $p_j \geq j\alpha/2m$. Under the approximation specified above we get

$$\begin{aligned} F_j(a^{\triangle}_{\alpha,j}) &= \frac{\alpha}{2m\mathbb{E}\frac{1}{1+\tilde{N}_j}} \\ &\leq \frac{\alpha}{2m}(1+\mathbb{E}\tilde{N}_j) \end{aligned}$$

by Jensen's inequality. Re-writing this we get

$$\begin{aligned} F_j(a^{\triangle}_{\alpha,j}) &\leq \frac{\alpha}{2m}\left(1+\sum_{k\neq j} 2F_j(a^{\triangle}_{\alpha,k})\right) \\ &= \frac{\alpha}{2m}\left(1+\sum_{k<j} 2F_j(a^{\triangle}_{\alpha,k}) + \sum_{k>j} 2F_k(a^{\triangle}_{\alpha,k})\right) \end{aligned}$$

The two leftmost bracketed terms are jointly bounded above by j, hence

$$F_j(a^{\triangle}_{\alpha,j}) \leq \frac{\alpha}{2m}\left(2j + 2\sum_{k>j} F_k(a^{\triangle}_{\alpha,k})\right)$$

But, by assumption, $F_j(\theta_{0j}) \geq j\alpha/m$ and so

$$F_j(a^{\triangle}_{\alpha,j}) - F_j(\theta_{0j}) \leq \frac{\alpha}{2m}\left(\sum_{k>j} 2F_k(a^{\triangle}_{\alpha,k})\right)$$

The right hand side term is positive, hence $F_j(a^{\triangle}_j) \leq F_j(\theta_{0j})$ and so θ_{0j} must lie inside the approximation to the Bayes rule.

Therefore, for all $p_j \geq \alpha j/2m$, we can declare that $\theta_{0j} \in [a^{\triangle}_{\alpha,j}, b^{\triangle}_{\alpha,j}]$. With induction, this defines a step-down procedure; modulo the stated approximation we can continue declaring $\theta_{0j} \in [a^{\triangle}_{\alpha,j}, b^{\triangle}_{\alpha,j}]$ until some j violates $p_j \geq j\alpha/2m$, the well-known 'Simes line' (Simes, 1986).

Maximizing the length of our 'run' of steps $j+1$ to j is achieved by ordering the p^*-values, so $p^*_{[1]} \leq p^*_{[2]} \leq ... \leq p^*_{[m]}$. The procedure defined in this way is therefore to search for

$$\max\{j : p^*_{[j]} \leq j\alpha/m\}$$

This is the very well-known Benjamini-Hochberg (BH) algorithm (Benjamini and Hochberg, 1995), applied to p^*-values for the θ_j which are by assumption independent in the posterior. Applied to p-values from independent test statistics, the BH algorithm control the

expected FDR at a pre-specified level, where the expectation is in the frequentist sense of infinite numbers of replications. In the decision-theoretic setting, BH represents an attractively simple approximate step-down method to answer whether elements of θ_{0j} are included in the Bayes intervals $[a^{\triangle}_{\alpha,j}, b^{\triangle}_{\alpha,j}]$.

6.2 Appendix 2

Taking the expectation with respect to the posterior, we find

$$\begin{aligned}\mathbb{E}L_\gamma &= \gamma tr(R) + tr\left(R^{-1}\mathbb{E}(\theta - d)(\theta - d)^T\right) \\ &= \gamma tr(R) + tr\left(R^{-1}(\bar{\theta} - d)(\bar{\theta} - d)^T + R^{-1}\mathbb{E}(\theta - \bar{\theta})(\theta - \bar{\theta})^T\right),\end{aligned}$$

where $\bar{\theta}$ denotes the posterior mean of θ. Now, re-arranging the term in d as above, it is easily seen to be non-negative, and is thus minimized (to zero) when d is the posterior mean.

It remains to minimize the expected loss with respect to R. As a convenient shorthand, we write $V = \mathbb{E}(\theta - \bar{\theta})(\theta - \bar{\theta})^T$, the posterior variance. Re-parameterizing the decision problem as $R = \gamma^{-1/2}V^{1/2}A$, where $V^{1/2}$ is the unique semi-definite square root of V and A is positive definite, we must minimize

$$\begin{aligned}\mathbb{E}L_\gamma &= \gamma^{1/2}tr(V^{1/2}A) + \gamma^{1/2}tr(A^{-1}V^{1/2}) \\ &= \gamma^{1/2}tr\left(V^{1/2}(A + A^{-1})\right)\end{aligned}$$

By considering the inequality

$$tr\left((A^{1/2} - A^{-1/2})V^{1/2}(A^{1/2} - A^{-1/2})\right) \geq 0$$

we observe that the expected loss must obey

$$\gamma^{1/2}tr\left(V^{1/2}(A + A^{-1})\right) \geq \gamma^{1/2}2\, tr\left(V^{1/2}\right),$$

this minimum being attained when we choose A to be the identity matrix. Hence we find that the Bayes rule simply sets

$$d = \mathbb{E}\theta|Y, R^2 = \gamma^{-1}\mathrm{Cov}\theta|Y.$$

6.3 Appendix 3

The loss is

$$\begin{aligned} L_\gamma &= \gamma tr(R) + \mathbb{E}_{Y^*|\theta}(Y^* - Xd)^T \Pi(X)^T R^{-1} \Pi(X)(Y^* - Xd), \\ &= \gamma tr(R) + tr(R^{-1}\Pi(X) \left(\mathbb{E}_{Y^*|\theta}(Y^* - Xd)(Y^* - Xd)^T\right) \Pi(X)^T) \end{aligned}$$

Taking the expectation of the loss, the terms in $\mathbb{E}_{Y^*|\theta}$ are replaced by posterior predictive expectations. Minimizing the expected loss then proceeds in a similar way to Appendix 2. The expected loss can be decomposed;

$$\begin{aligned} \mathbb{E}L_\gamma &= \gamma tr(R) + tr(R^{-1}\Pi(X) \left(\mathcal{E}(Y^* - X\Pi(X)\bar{Y})(Y^* - X\Pi(X)\bar{Y})^T\right) \Pi(X)^T) \\ &\quad + tr(R^{-1}\Pi(X) \left(\mathcal{E}(X\Pi(X)\bar{Y} - Xd)(X\Pi(X)\bar{Y} - Xd)^T\right) \Pi(X)^T), \end{aligned}$$

where $\bar{Y}$ denotes the mean of the posterior predictive. The term in d is minimized to zero by setting $d = \Pi(X)\bar{Y}$. In the term involving Y^*, the exterior $\Pi(X)$ terms may be brought inside the expectation terms. With the cancellation that follows, and minimization with respect to R as before, we find that the Bayes rule sets

$$\{d, R^2\} = \{\Pi(X)\bar{Y}, \Pi(X) \left(\mathcal{E}(Y^* - \bar{Y})(Y^* - \bar{Y})^T\right) \Pi(X)^T)\}.$$

Now, under the assumption (in the predictive distribution) that the Y_i^* are independent, the Bayes rule for R^2 can be written as

$$R^2 = (X^T X)^{-1} X \text{diag}\{\text{Var} Y_i^*\} X^T (X^T X)^{-1}.$$

This parallels the familiar 'sandwich' forms used in the Estimating Equations paradigm.

We note that, if the classical linear model is assumed as the data-generating mechanism, with homoskedastic Normal error terms and non-informative priors, then the Bayes rule for the intervals reduces to the standard likelihood-based confidence intervals. At the other extreme, for sufficiently flexible models and non-informative priors, the Bayes rules reproduce frequentist sandwich-based intervals (for fixed X), and conseqently intervals will have asymptotically correct frequentist coverage.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence (C/R: P123-133, 135-139, 1201-1201). *Journal of the American Statistical Association*, 82:112–122.

Bland, M. (2000). *An introduction to medical statistics, 3rd Edition.* Oxford University Press.

Brown, L. D. (2000). An essay on statistical decision theory. *Journal of the American Statistical Association*, 95(452):1277–1281.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (C/R: P123-135). *Journal of the American Statistical Association*, 82:106–111.

Casella, G., Hwang, J. T. G., and Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica*, 3:141–155.

Casella, G., Hwang, J. T. G., and Robert, C. P. (1994). Loss functions for set estimation. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics, V*, pages 237–251. Springer-Verlag Inc.

Caudill, S. B. (1988). Type I errors after preliminary tests for heteroscedasticity. *The Statistician: Journal of the Institute of Statisticians*, 37:65–68.

Christensen, R. (2005). Testing fisher, neyman, pearson, and bayes. *The American Statistician*, 59(2):121–126.

Cohen, A. and Strawderman, W. E. (1973). Admissible confidence interval and point estimation for translation or scale parameters. *The Annals of Statistics*, 1:545–550.

DeGroot, M. H. (2004). *Optimal Statistical Decisions*. Wiley-IEEE.

Dixon, D. O. (1976). Interval estimates derived from Bayes testing rules. *Journal of the American Statistical Association*, 71:406–408.

Dixon, D. O. and Duncan, D. B. (1975). Minimum Bayes risk t-intervals for multiple comparisons. *Journal of the American Statistical Association*, 70:822–831.

Duncan, D. B. (1975). t tests and intervals for comparisons suggested by the data. *Biometrics*, 31:339–359.

Efron, B. (2008). Microarrays, empirical bayes, and the two-groups model. *Statistical Science*. in press.

Fearn, T. (2007). Discussion of FDR and Bayesian Decision Rules, by Müller Parmigiani and Rice. In *Proceedings of the Valencia/ISBA 8th world meeting on Bayesian Statistics*. Oxford University Press.

Forsberg, L. and Kristiansson, U. (1999). On type I errors after a preliminary test for heteroscedasticity. *Journal of the Royal Statistical Society, Series D: The Statistician*, 48:63–72.

Freedman, D. (1999). Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302.

Gupta, S. and J.O., B., editors (1993). *Statistical Decision Theory and Related Topics V*. Springer-Verlag, New York.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Symp. Math. Statist. and Probability*, volume 2, pages 221–233. Univ. of California Press.

Inoue, L. Y., Berry, D. A., and Parmigiani, G. (2005). Relationship between Bayesian and frequentist sample size determination. *The American Statistician*, 59(1):79–87.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192.

Lindley, D. V. (1997). The choice of sample size (Disc: P139-166). *The Statistician: Journal of the Institute of Statisticians*, 46:129–138.

McCullagh, P. and Nelder, J. A. (1999). *Generalized Linear Models (Second Edition)*. Chapman & Hall Ltd.

Moreno, E. and Girón, F. J. (2006). On the frequentist and bayesian approaches to hypothesis testing. *Statistics & Operations Research Transactions*, pages 3–28.

Mueller, P., Giovanni, P., and Rice, K. (2007). Fdr and bayesian multiple comparisons rules. In *Proceedings of the Valencia/ISBA 8th World Meeting on Bayesian Statistics*. Oxford University Press.

Robert, C. P. (1994). *The Bayesian Choice: a Decision-theoretic Motivation*. Springer-Verlag Inc.

Robert, C. P. (2001). *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*. Springer-Verlag Inc.

Robert, C. P. and Casella, G. (1994). Distance weighted losses for testing and confidence set evaluation. *Test*, 3(1):163–182.

Robins, J. and Wasserman, L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association*, 95(452):1340–1346.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons.

Schervish, M. (1995). *Theory of Statistics*. Springer-Verlag.

Senn, S. (2000). Consensus and controversy in pharmaceutical statistics (Pkg: P135-176). *Journal of the Royal Statistical Society, Series D: The Statistician*, 49(2):135–156.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754.

Szpiro, A. A., Rice, K., and Lumley, T. (2007). Model-robust bayesian regression and the sandwich estimator. Technical Report 320, University of Washington Biostatistics Working Paper Series, http://www.bepress.com/uwbiostat/paper320.

Van Belle, G., Fisher, L., Heagerty, P., and Lumley, T. (2004). *Biostatistics; A Methodology For The Health Sciences (2nd Edition)*. Wiley, Hoboken, New Jersey.

Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–830.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67:187–191.

Young, G. and Smith, R. (2005). *Essentials of Statistical Inference*. Cambridge University Press, Cambridge.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zeger, S. L., Liang, K.-Y., and Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72:31–38.